PROJECT ON PREDICTION OF SLEEP DISORDER

WEST BENGAL STATE UNIVERSITY

SUBJECT: STATISTICS

ROLL:6242125 NO:17583

REG NO.: 1082121400105

PAPER CODE: STSADSE06P

- SUMMARY:

This project is prediction of sleep disorder present in an individual.we can summarize the whole project as follows:

- First there is a brief introduction about the types of sleep disorder and their description.
- The second part contains brief description about the data on which we will work.
- The third part consists data preprocessing.
- The fourth part is the analysis of the data by using various diagrams.
- The fifth part consists the fitting of suitable model to the data.
- The sixth part is the interpretation of the output about the model we have fitted.
- The next part is the prediction of the response variable and check the accuracy of the model.
- The next part contains the distribution of the importance of different variables on the response variable.
- The last section respectively conclude the project and list the references of data used in the making the project

- ## INTRODUCTION:

Sleep disorders are a range of conditions that impact an individual's ability to sleep well on a regular basis. These disorders disrupt the normal sleep pattern, affecting the quality, timing, and duration of sleep. The resulting sleep disturbances can have significant effects on daily functioning, overall health, and quality of life. Insomnia and Sleep Apnea are two most common sleep disorder in common days.

Sleep apnea is a sleep disorder characterized by repeated interruptions in breathing during sleep. These disruptions can last from a few seconds to a minute or more and can occur multiple times per hour. The condition results in reduced oxygen levels in the blood and fragmented sleep, leading to daytime fatigue and other health issues.

Insomnia is a sleep disorder characterized by difficulty falling asleep, staying asleep, or achieving restorative sleep, despite having the opportunity to sleep. It results in impaired daytime functioning, such as fatigue, mood disturbances, and difficulty concentrating. Insomnia can be acute (short-term) or chronic (long-term), with varying degrees of severity.

OBJECTIVE:

The objective of the project is to prediction of the sleep disorder of an person based on some information of the independent variables.As there are so many variables present as predictor we want to know how the variables effect the response i.e. sleep disorder. We also want to find the importance of each of the variables for prediction.

- Experiment and Result
- Independent and Dependent Variables:

The dataset “Sleep Health and Lifestyle Dataset” used for this project comes from Kaggle.com. Table 1 shows the names of the columns in the dataset and their explanations [13]. Most columns in the dataset are rated subjectively, such as quality of sleep and stress level.

- Data source:

https://www.kaggle.com/code/uurdemirkaya/sleep-health-and-lifestyle-dataset-analysis

Table-1: Dataset Columns

Variables	Explanation
1. Person ID	A number for each participant
2. Gender	The gender of the person (Male/Female)
3. Age	The age of the person
4. Occupation	The occupation of the person
5. Sleep Duration	The number of hours the person sleeps per day
6.Quality of Sleep	A subjective rating of the quality of sleep, ranging from 1 to 10
7.Physical Activity Level	The number of minutes the person engages in physical activity daily
8.Stress Level	A subjective rating of the stress level experienced by the person, ranging from 1 to 10
9.BMI Category	The BMI category of the person (Normal, Normal Weight, Overweight, Obese)
10.Blood Pressure	The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure
11. Heart Rate	The resting heart rate of the person in beats per minute
12.Daily Steps	The number of steps the person takes per day
13. Sleep Disorder	The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea)

- The **"Sleep Disorder"** is the dependent variable, that is the target that is trying to predict. The other columns are the independent variables. In total, there are 559 rows and 13 columns in the dataset.

- Data Preprocessing :

In the beginning, the dataset needs to be preprocessed. First, columns that are not consistent need to be changed. For instance, the “Blood Pressure” column. The highest blood pressure and lowest were together. In this way, it is difficult to see and classify it. The first step was to split the "Blood Pressure" column into separate ones. The “Blood Pressure” column was dropped, and in its place were “sys pressure” and “dias pressure” for each person and then a new variable is formed ,

MAP = [sys pressure+(2x dias pressure)]/3.

Then the data was split with 70% used for training and 30% used for testing. The training data was fitted into the models. Then, the accuracy of the models was calculated. Using the “predict” method, and putting the test data as input, the target variable, “Sleep Disorder” was predicted.

- ## Analysis:

Before predicting, a series of analyses were done on the data to gain more insight. This helps have a better understanding of the dataset.

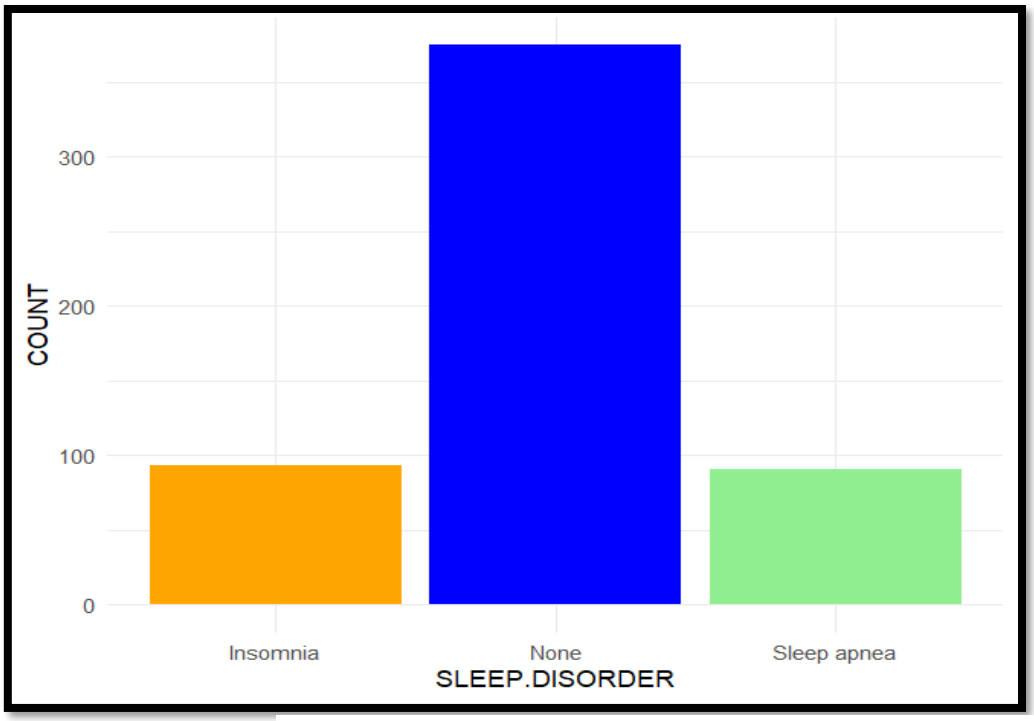

Fig-1: Distribution of Sleep Disorder

For the column “Sleep Disorder”, there are three conditions: None, Sleep Apnea, or Insomnia. Sleep Apnea is a kind of sleep disorder in which an individual suffers from breathing problems during sleep. Insomnia is difficulty falling asleep or staying asleep. As shown in Figure 1 is the distribution of sleep disorder. These three types of conditions are displayed on the x-axis, and the y-axis shows the number. The color blue is for people with no sleep disorder, orange is for people with sleep apnea, and green is for people with insomnia. Figure 1 shows that more than 350 participants are in the “none” category. As for the people with sleep disorders, there is an almost equal number of people with sleep apnea and insomnia. The number is about 70.

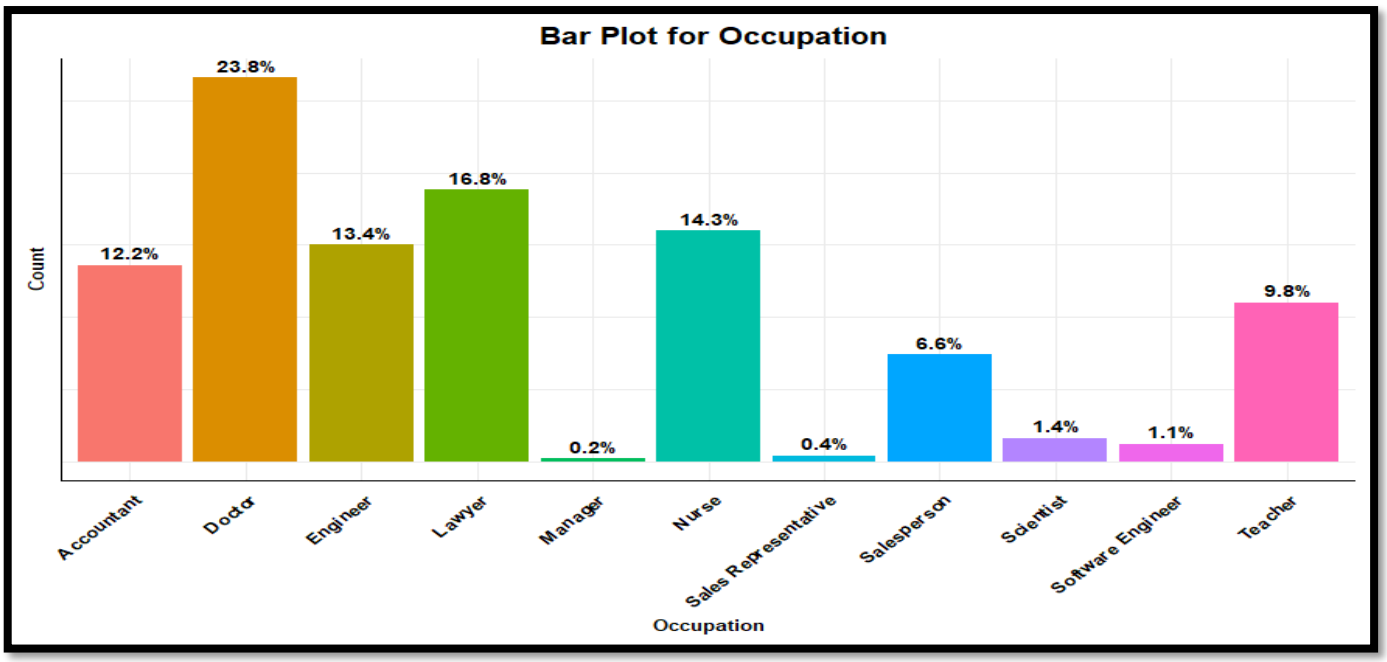

Figure 2: Distribution of occupation

Individuals in the data are from different occupations There are total 11 occupations like Doctor,Engineer etc. From the above bar diagram of occupation we get an idea about the distribution of occupation in the data. Here the figure shows that most of the people in the data are doctors and least no. of people are from the occupation manager.

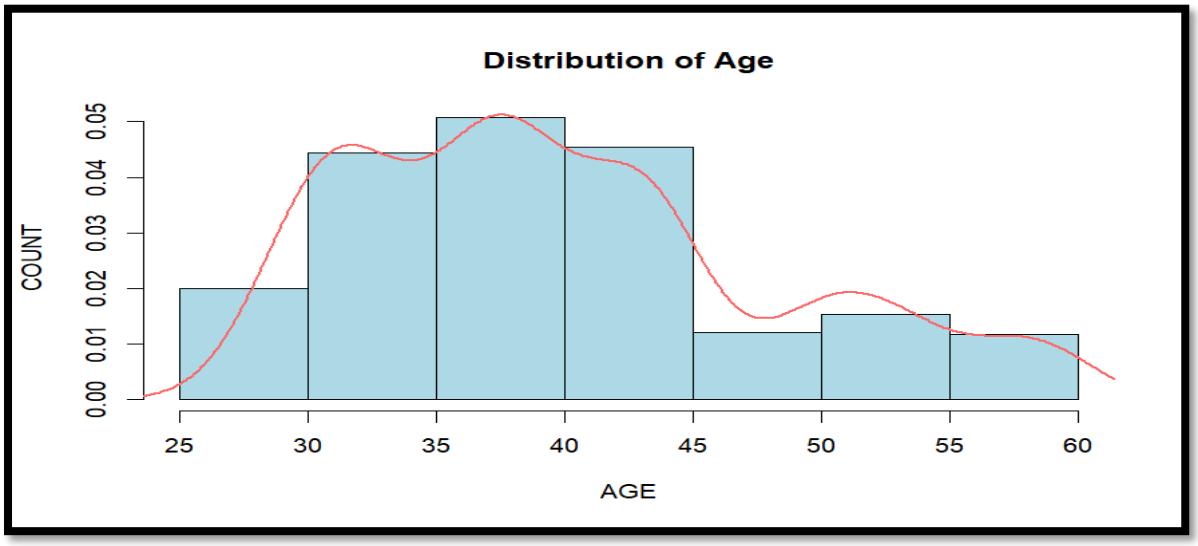

Figure 3: Distribution of age

Like people of different occupation the data is of people of different ages. We can clearly get an idea about the distribution of age in the data from the above bar diagram. Here we see that maximum people in the data set are of the age between 30-45 years.

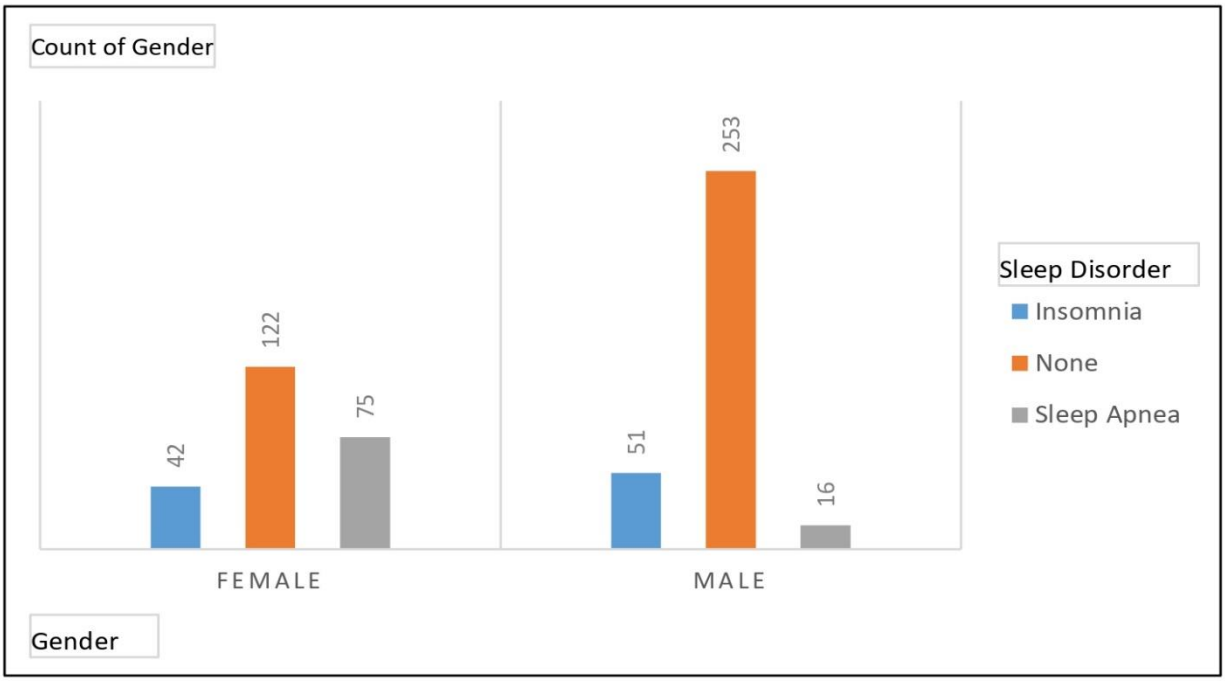

Figure 4: Sleep disorder vs. gender

- Sleep Disorder and Gender:

There is a big difference in gender. As shown in Figure 3, male and female each has three bars, which indicates their own distribution of sleep disorder. Like Figure 1, orange is for no sleep disorder, blue is for insomnia, and grey is for sleep apnea. On the left is the distribution for females and the right is for males. The number of males that have no sleep disorder is 253, which is twice as much as females. In addition, less than 16 males have sleep apnea and around 51 males have insomnia, which is almost the same as females. However, there are more females with sleep apnea, 75 females.

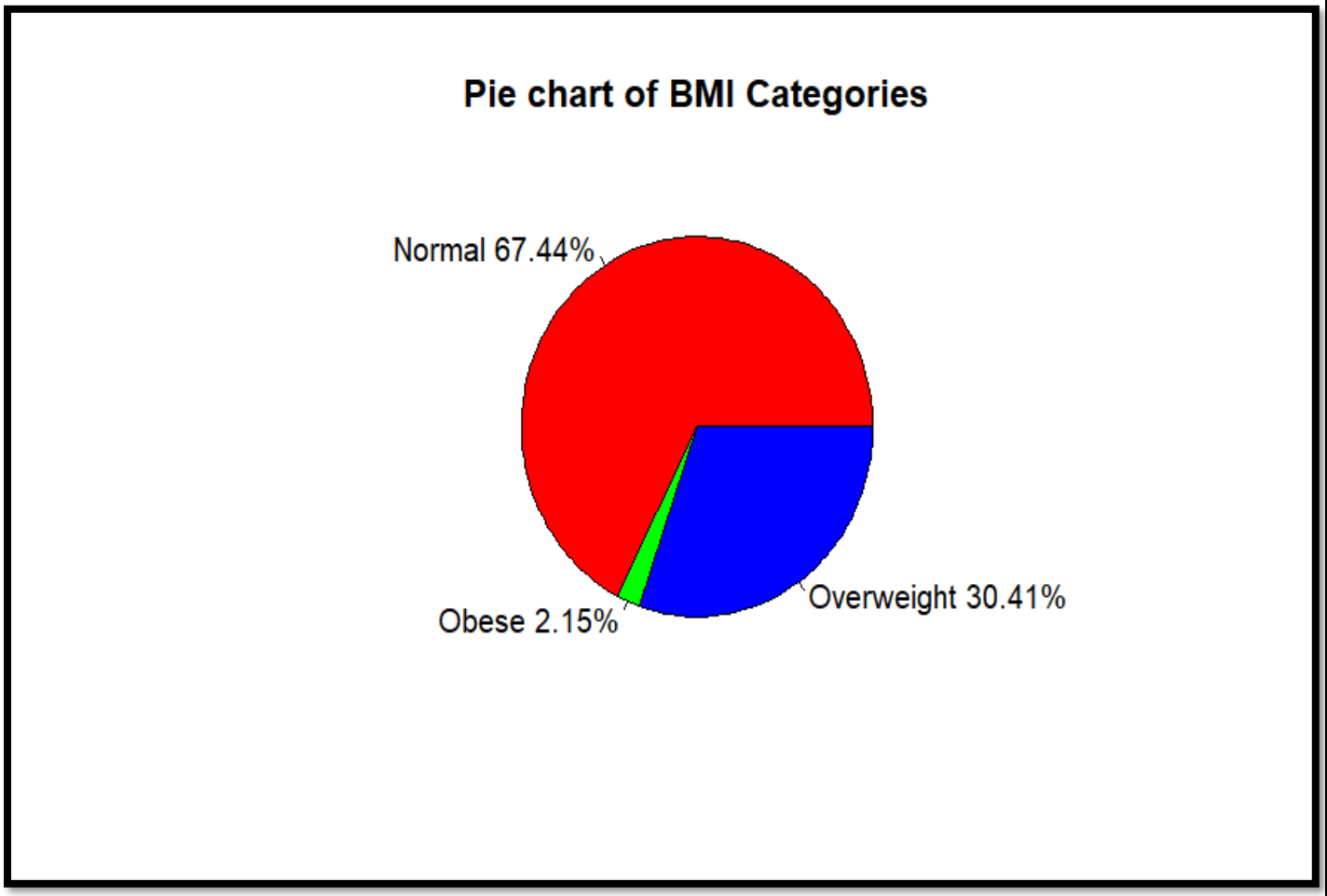

Figure 5:pie chart of BMI category

- ## Sleep Disorder and BMI Category:

  Sleep disorders not only can be affected by gender but also by BMI (body mass index) category. There are four parts in the "BMI category" column: overweight, normal, obese, and normal weight. Here we construct a pie chart showing the percentage of person in different BMI categories. From the figure we can observe that there are 67.44% individuals in Normal BMI category,30.41% individuals are overweight and 2.15% are obese.

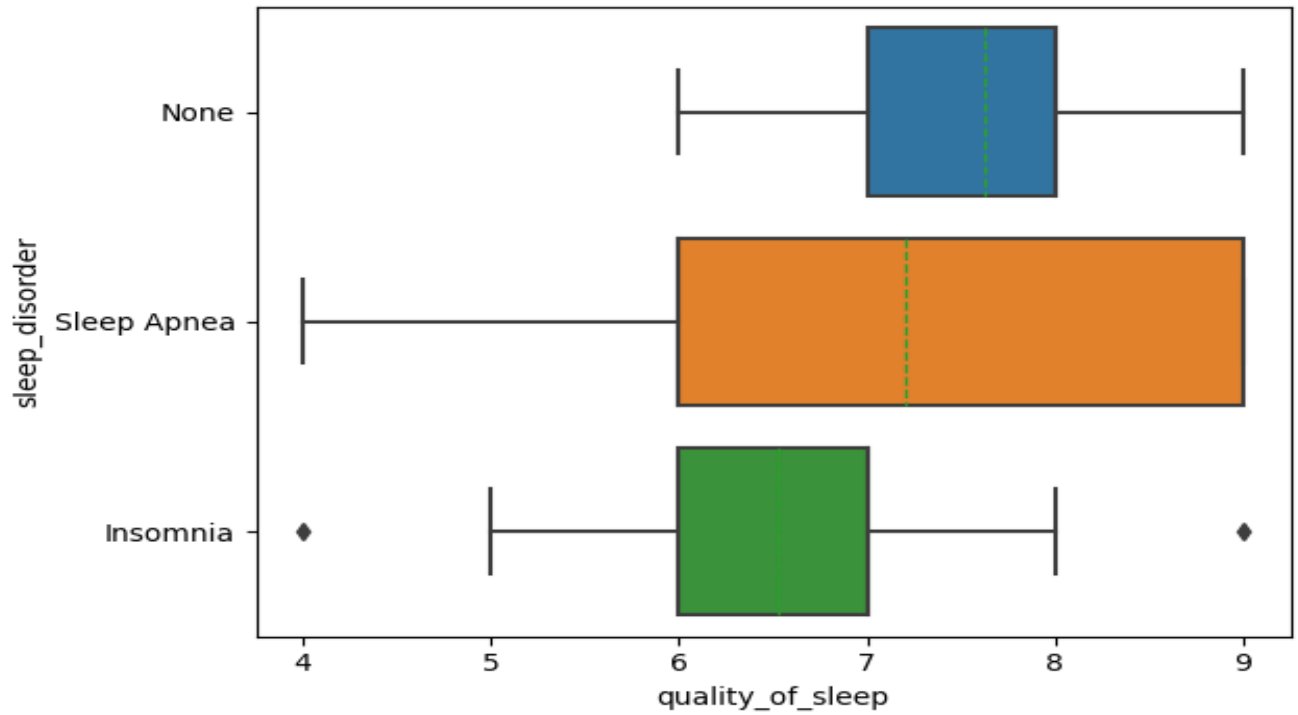

Figure 6: box plot of quality of sleep

This box plot shows the distribution of sleep quality scores across three sleep disorder categories: None, Sleep Apnea, and Insomnia.

- None has higher median sleep quality,Smaller interquartile range (IQR), indicating less variability in sleep quality and Symmetrical whiskers, suggesting balanced distribution.
- Sleep Apnea has Moderate median sleep quality, slightly lower than None.Larger IQR, indicating more variability and Whiskers extend further, suggesting some lower and higher sleep quality scores.
- Insomnia Lowest median sleep quality,Small IQR, indicating less variability. Presence of outliers, suggesting some higher sleep quality scores than typical.

The plot highlights that individuals with no sleep disorder generally have higher sleep quality, while those with insomnia report the lowest quality.

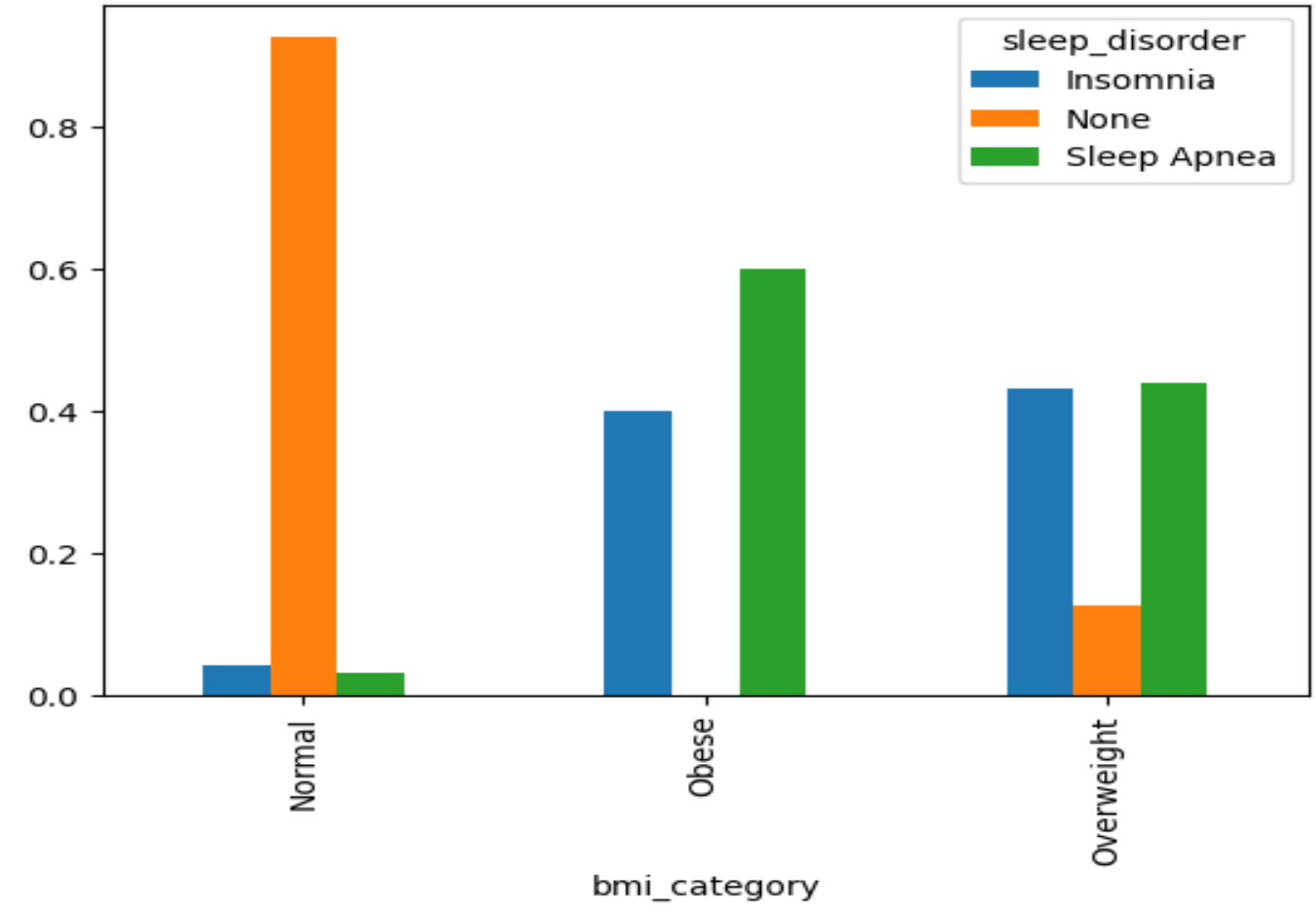

Figure 7: distriibution of sleep disorder for BMI category

This bar chart shows the distribution of sleep disorders across different BMI categories: Normal, Obese, and Overweight.

- Normal BMI: has Predominantly no sleep disorder..
- Obese BMI:has High prevalence of sleep apnea.,Moderate presence of insomnia.and Almost no cases with no sleep disorder.
- Overweight BMI has Balanced distribution between insomnia and sleep apnea.and Few cases with no sleep disorder.

The chart suggests that individuals with normal BMI are more likely to have no sleep disorder, while those who are obese are more likely to experience sleep apnea. Overweight individuals show a mixed pattern between insomnia and sleep apnea.

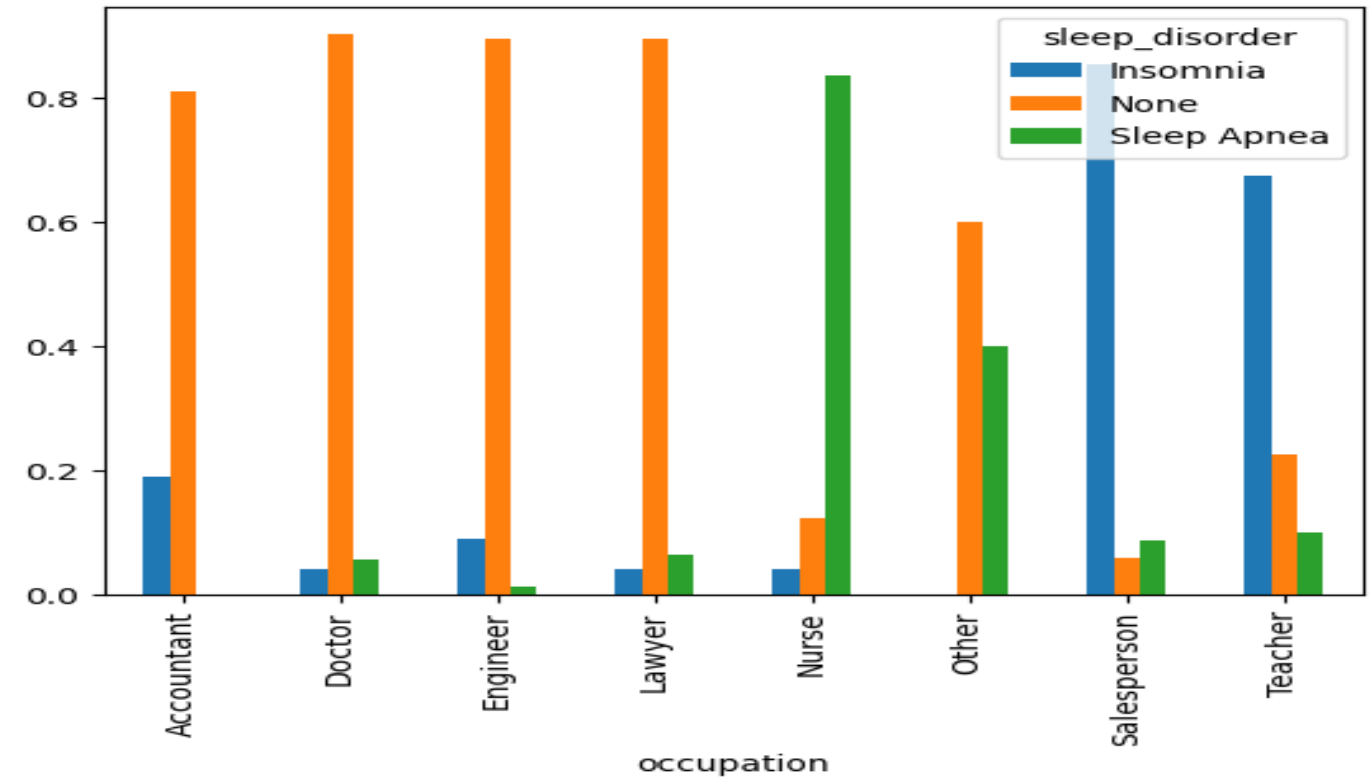

Figure8: Distribution sleep disoreder among different occupation

This bar chart shows the distribution of sleep disorders across various occupations.

- Accountant, Doctor, Engineer, Lawyer:
  - High prevalence of no sleep disorder and minimal presence of insomnia and sleep apnea.
- Nurse:
  - Significant prevalence of sleep apnea,Some cases with no sleep disorder and Minimal insomnia.
- Other:
  - Balanced distribution across all categories and Higher presence of sleep apnea.
- Salesperson:
  - High prevalence of insomnia, Moderate sleep apnea presence and Fewer cases with no sleep disorder.
- Teacher:
  - High prevalence of insomnia and Some cases with no sleep disorder and sleep apnea.

The chart indicates that more physically demanding or stressful jobs like nursing and sales may have higher sleep disorder rates. Conversely, more sedentary or routine jobs like accounting and engineering show fewer sleep disorders.

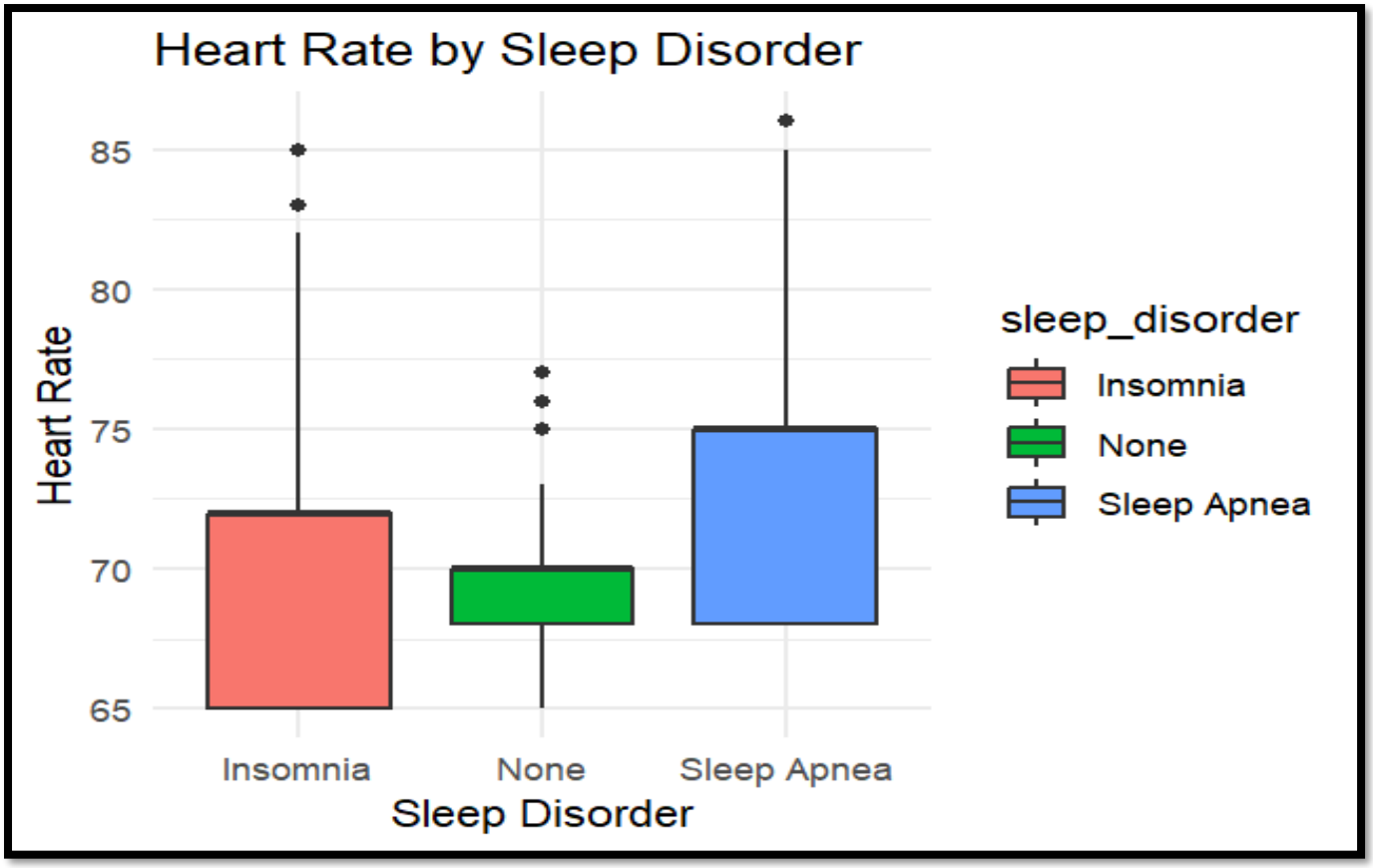

Figure 9: box plot of heart rate

The image presents a boxplot visualizing the relationship between heart rate and different sleep disorders: Insomnia, None and Sleep Apnea.

Sleep Apnea has the highest median heart rate.

Insomnia and None have similar median heart rates, with None being slightly lower.

Sleep Apnea has the widest IQR, suggesting more variability in heart rates within this group.

Insomnia and None have narrower IQRs, indicating less variability.
Sleep Apnea and None have a few outliers with higher heart rates.

Insomnia has no outliers.

- ## Model fitting :

There are 3 categories of response variable Y viz. Insomnia, None and Sleep Apnea with probabilities π_1 , π_2, π_3 . $\sum_j \pi_j = 1$.
Here the response variable Y denotes the Sleep disorder which has 3 categories viz. Insomnia, Sleep Apnea and None and there are 11 independent or predictor variables among them some are categorical some of them are continuous variable.

As the response variable is categorical with 3 categories, we will fit a multiple logistic regression model using 11 independent variables.

Let the independent variables are denoted by

X_1 =Gender of the individual
X_2 =Age of the individual
X_3 =Occupation
X_4 =Sleep duration
X_5 =Quality of Sleep
X_6 =Physical activity
X_7 =Stress level
X_8 =BMI category
X_9 = Heart rate
X_{10} =Daily steps
X_{11} =MAP

With 'Sleep Apnea' as the baseline category for Y, the model is:

$$\log(\pi_j/\ \pi_3) = \alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2 + \ldots\ldots\ldots\ldots + \beta_{j11}x_{11}\ ,\ \ j=1,2$$

- ## *R codes used to fit the model:*

```
library(VGAM)
df<-read.csv("projecT (1).csv",header=TRUE)

#Partitioning the data
library(caret)
index <- createDataPartition(dataset$Sleep.Disorder, p = .70, list = FALSE)
train <- dataset[index,]
test <- dataset[-index,]

#Fitting the multinomial model
fit<-vglm(Sleep.Disorder~Gender+Age+Occupation+Sleep.Duration+Quality.of.Sleep+
Physical.Activity.Level+Stress.Level+BMI.Category+Heart.Rate+Daily.Steps+MAP,family="multinomial",
data=train, weights=NULL,trace=TRUE)
coef(fit,matrix=TRUE)
```

- Output:

```
Coefficients:
                                    Estimate   Std. Error z value Pr(>|z|)
(Intercept):1                     -2.350e+01  2.277e+03      NA       NA
(Intercept):2                      1.765e+02  2.276e+03      NA       NA
GenderMale:1                      -1.239e+00  4.417e+00  -0.280   0.7791
GenderMale:2                      -1.338e+00  3.306e+00  -0.405   0.6856
Age:1                              3.739e-01  4.852e-01   0.771   0.4410
Age:2                             -1.826e-01  2.703e-01  -0.675   0.4994
OccupationDoctor:1                -5.954e+01  2.275e+03  -0.026   0.9791
OccupationDoctor:2                -9.328e+00  2.275e+03  -0.004   0.9967
OccupationEngineer:1              -3.543e+01  2.275e+03  -0.016   0.9876
OccupationEngineer:2              -5.466e+00  2.275e+03  -0.002   0.9981
OccupationLawyer:1                -3.407e+01  2.275e+03  -0.015   0.9881
OccupationLawyer:2                -3.145e+00  2.275e+03  -0.001   0.9989
OccupationManager:1               -2.946e+01  3.368e+04      NA       NA
OccupationManager:2                3.438e+00  2.387e+04   0.000   0.9999
OccupationNurse:1                 -5.883e+01  2.275e+03  -0.026   0.9794
OccupationNurse:2                 -1.495e+01  2.275e+03  -0.007   0.9948
OccupationSales Representative:1  -7.983e+01  1.695e+04      NA       NA
OccupationSales Representative:2  -3.960e+00  1.803e+04   0.000   0.9998
OccupationSalesperson:1           -2.762e+01  4.082e+03  -0.007   0.9946
OccupationSalesperson:2           -7.658e-01  4.082e+03   0.000   0.9999
OccupationScientist:1             -7.291e+01  6.828e+03      NA       NA
OccupationScientist:2             -1.475e+01  2.275e+03  -0.006   0.9948
OccupationSoftware Engineer:1     -3.799e+01  1.900e+04  -0.002   0.9984
OccupationSoftware Engineer:2     -1.141e+00  9.911e+03   0.000   0.9999
OccupationTeacher:1               -4.806e+01  2.275e+03  -0.021   0.9831
OccupationTeacher:2               -1.877e+01  2.275e+03  -0.008   0.9934
Sleep.Duration:1                  -7.098e+00  4.141e+00      NA       NA
Sleep.Duration:2                  -4.093e+00  2.601e+00  -1.574   0.1156
Quality.of.Sleep:1                -1.816e+00  3.671e+00      NA       NA
Quality.of.Sleep:2                 1.186e+00  3.157e+00   0.376   0.7073
Physical.Activity.Level:1          9.438e-02  9.490e-02   0.995   0.3200
Physical.Activity.Level:2          9.927e-02  7.539e-02      NA       NA
Stress.Level:1                     3.921e+00  4.629e+00   0.847   0.3970
Stress.Level:2                     6.693e-01  2.389e+00   0.280   0.7793
BMI.CategoryNormal Weight:1       -9.674e+00  1.433e+01  -0.675   0.4996
BMI.CategoryNormal Weight:2        9.385e+00  6.188e+00      NA       NA
BMI.CategoryObese:1               -2.796e+01  1.876e+01      NA       NA
BMI.CategoryObese:2               -9.148e+00  6.142e+03      NA       NA
BMI.CategoryOverweight:1          -1.591e+01  1.243e+01  -1.280   0.2005
BMI.CategoryOverweight:2           1.118e+01  5.161e+00   2.167   0.0303 *
Heart.Rate:1                      -1.193e+00  9.787e-01      NA       NA
Heart.Rate:2                      -1.079e+00  5.002e-01  -2.157   0.0310 *
Daily.Steps:1                     -6.378e-03  3.873e-03      NA       NA
Daily.Steps:2                     -6.016e-04  6.565e-04  -0.916   0.3595
MAP:1                              2.216e+00  1.590e+00   1.394   0.1634
MAP:2                             -6.889e-01  3.249e-01  -2.120   0.0340 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Residual deviance: 167.2214 on 740 degrees of freedom

Log-likelihood: -83.6107 on 740 degrees of freedom

Number of Fisher scoring iterations: 19
```

- # INTERPRETATION OF THE OUTPUT:

- ## Intercepts:

In the output, two intercepts are provided, likely corresponding to two categories compared to a baseline (third category i.e sleep apnea).

Intercepts Represent the baseline log odds for each category compared to the reference category.

Large standard errors and NA for significance indicate instability or redundancy.

- Intercept for Category 1:
- Estimate: −23.5
- Interpretation: The log odds of being in Category 1(Insomnia) compared to Category 3 (baseline i.e. Sleep Apnea) when all predictors are zero is 13.91. Exponentiating this gives the odds ratio: $\text{Odds Ratio}_{\text{Category 1}} = e^{-23.5} \approx 7.11 \times 10^{-11}$.
  This odds ratio indicates that Insomnia extremely unlikely to occur compared to Sleep apnea.

- Intercept for Category 2:
- Estimate: 176.5
- Interpretation: The log odds of being in Category 2(None) compared to Category 3 (baseline i.e. Sleep Apnea) when all predictors are zero is 176.5. Exponentiating this gives: $\text{Odds Ratio}_{\text{Category 2}} = e^{176.5} \approx 1.398 \times 1076$.
- ## Coefficients:

The coefficients represent the change in the log odds of the dependent variable for a one-unit increase in the predictor variable, holding other variables constant.

- GenderMale:1:
- Estimate: −1.239
- Interpretation: Presence of Gender Male variable decreases the log odds of being in Category 1(Insomnia) compared to Category 3 (baseline i.e. Sleep Apnea) by 1.239, assuming all other variables are held constant.

- Standard Error:

The standard error (SE) provides a measure of the variability or precision of the coefficient estimates. It indicates how much the estimate might vary from sample to sample. A smaller standard error suggests more precise estimates, while a larger standard error indicates greater uncertainty.

  - GenderMale:1
  - Estimate: -1.239
  - Standard Error (SE): 4.417
  - The standard error (4.417) tells us about the variability of the coefficient estimate (-1.239). Specifically, it suggests that the estimate might vary by ±4.417 units in different samples.

- Z value:

The z-value (or z-score) in the context of a logistic regression model is used to determine whether a coefficient is significantly different from zero. It is calculated by dividing the coefficient estimate by its standard error:

$$\text{z-value} = \text{Estimate} \div \text{Standard Error}$$

The z-value helps in testing the null hypothesis that the coefficient is equal to zero (i.e., it has no effect). A large absolute z-value indicates that the coefficient is significantly different from zero.

- High Absolute Z-value: A high absolute z-value (typically greater than ±1.96) indicates that the coefficient is significantly different from zero. This suggests that the predictor has a statistically significant effect on the outcome variable relative to the reference category.
- Low Absolute Z-value: A low absolute z-value indicates that the coefficient is not significantly different from zero. This suggests that the predictor may not have a significant effect.

- A z-value for Occupation Doctor:1 −0.026 suggests that the parameter estimate is not significantly different from zero. This implies that there is little evidence to suggest a meaningful effect of being in the Occupation Doctor category on occurring of sleep apnea compared to the occurring of Insomnia.
- The z-value for Occupation Doctor:2 = −0.004 is extremely close to zero..This imply that the occupation doctor has no meaningful effect on the occurance of no sleep disorder compared to occurance of the Sleep Apnea.

- P-value:

The p-value is the probability of observing a test statistic at least as extreme as the one actually observed in sample, under the assumption that the null hypothesis is true.

In the context of multinomial logistic regression, it assesses whether the coefficients of the predictor variables are significantly different from zero for each comparison of outcome categories.

- P-Value < 0.05: The predictor is statistically significant for that outcome category compared to the reference category.
- P-Value between 0.05 and 0.10: The predictor has marginal significance; it may be worth investigating further.
- P-Value > 0.10: The predictor is not statistically significant for that outcome category compared to the reference category.

- p-value of Sleep Duration

- Sleep.Duration:1:
- P-value = 0.0117.
- Sleep duration significantly affects the likelihood of being in Category 1 (Insomnia) compared to the reference category(Sleep Apnea)
- Sleep.Duration:2:
- P-value = 0.0228.
- Sleep duration significantly affects the likelihood of being in Category 2 (None)compared to the reference category(Sleep Apnea).

- PREDICTION:

We divided the whole dataset into two parts one was used for the model fitting called train data set and another was kept for prediction called test dataset.

First we will predict the response variable(Sleep Disorder) on the basis of the independent variables(Age, Gender,Heart rate etc.) for test dataset. In multiple logistic regression, "prediction" refers to the process of estimating the probability that a given observation belongs to a particular class or category based on the values of multiple predictor variables.

Applying prediction code in R we can get the probabilities of an individual belonging in the 3 categories(Insomnia, None and Sleep Apnea) for test data set. Comparing the probabilities we can conclude that an individual will belong to the category with highest probabilities. So we get the categories of the individuals of the test dataset.

We will predict the response variable for the test dataset using the fitted model based On the training dataset.

	Insomnia	None	Sleep Apnea
2	5.627943e-09	2.060358e-10	1.000000e+00
3	5.627943e-09	2.060358e-10	1.000000e+00
8	5.304055e-07	9.520547e-01	4.794477e-02
10	5.304055e-07	9.520547e-01	4.794477e-02
12	5.304055e-07	9.520547e-01	4.794477e-02
13	6.861554e-04	7.743249e-01	2.249889e-01
14	9.912653e-04	7.416495e-01	2.573592e-01
19	7.056210e-02	7.481314e-06	9.294304e-01

For observation 2,

The probability of occurance of Insomnia is 5.30×10^{-9}=0.000000005627943

The probability of occurance of no sleep disorder is 2.060×10^{-10} =0.0000000002060358

The probability of occurance of Sleep Apnea is 1.0000=1

This implies that the 2nd individual will have sleep Apnea.

```
predicted_sleepdisorder
[1] "Sleep Apnea"  "Insomnia" "None"  "None"       "None" "None"
[7] "None“         "Insomnia" "None" "None“        "None" "None"
[13] "None"          "None“  "None“  "None“        "None" "None“
[19] "None“          "None"  "None"  "None“        "None" "Sleep Apnea"
[25] "None“          "None"  "None“  "Sleep Apnea" "None" "None"
[31] "None“          "None"  "None"  "None“        "None" "None“
[37] "None“          "None"  "None " "None“        "None" "None"
```

By using “Predicted classes, we get which individual has which kind of sleep disorder.

Now to verify the model accuracy we can check whether the predicted categories of the individuals of the test set is accurate or not we form a confusion matrix. Here the confusion matrix looks like:

	Actual		
Predicted	Insomnia	None	Sleep Apnea
Insomnia	22	2	4
None	5	109	0
Sleep Apnea	0	1	23

Here the confusion matrix provides a detailed view of the classification results. Here's how to interpret the entries:

- Insomnia / Insomnia (22): The number of times the model correctly predicted Insomnia when it was actually Insomnia.
- Insomnia / None (2): The number of times the model predicted Insomnia when it was actually None.
- Insomnia / Sleep Apnea (4): The number of times the model predicted Insomnia when it was actually Sleep Apnea.

Similarly we can interpret the other components.

- Accuracy of the model:

From the confusion matrix we can interpret that the off diagonal elements of the matrix are misclassified. But the diagonal elements perfectly predicts the response using the fitted model. By using the elements of the confusion matrix we can get the accuracy of the fitted model.Here we get,

```
[1] "Accuracy: 0.9277
```

We can say that, An accuracy of 0.9277 indicates that your model correctly predicted the outcome for approximately 92.77% of the instances in your dataset which is the indication of the well-performing of the model.

- IMPORTANCE OF DIFFERENT PREDICTOR VARIABLES:

There are 12 different predictor variables to predict the sleep disorder of an individual. But it is natural to have different importance of different variables in predicting the response variable.

We can get idea about the importance of the different variables by forming a bar diagram of variables versus importance. This diagram also provides a clear and actionable view of how each feature affects the predictive model.

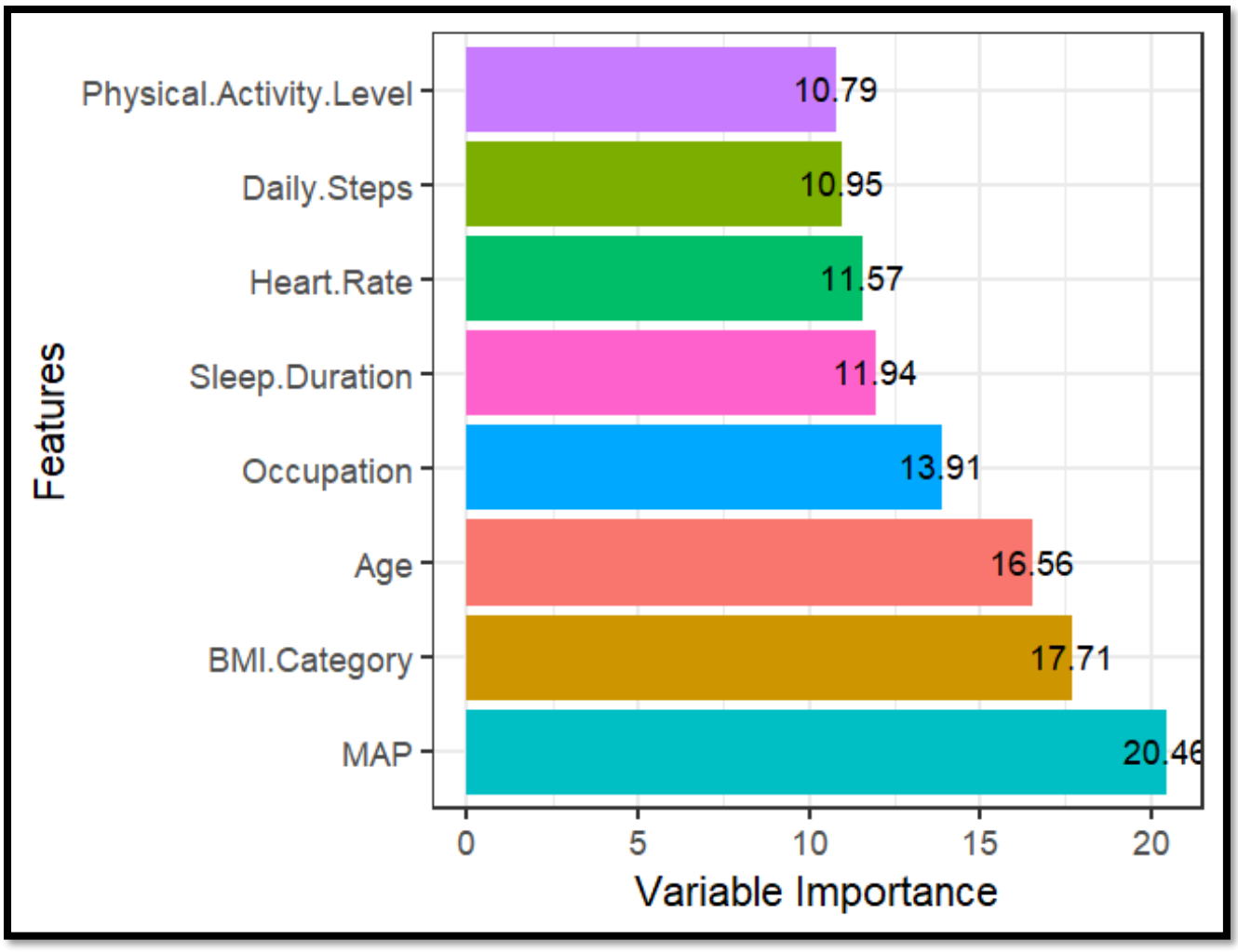

Figure: Bar diagram showing the importance of the selected features

The above diagram shows the importance of the independent variables to predict the reponse(sleep disorder).

CONCLUSION:

At the end of the project we can conclude that,

- This project aimed to develop a predictive model for sleep disorders using various features such as BMI category, age, and heart rate. Our model demonstrated an accuracy of 92.77%, indicating a high level of performance in distinguishing between individuals with and without sleep disorders. Key predictors identified include BMI category and heart rate, which were significant in predicting sleep disorders.
- The findings suggest that individuals with certain BMI categories and abnormal heart rates are at higher risk for sleep disorders, which has important implications for early diagnosis and intervention. By incorporating these predictors into healthcare practices, we can enhance early detection and personalized treatment plans for sleep disorders.

Overall, this project contributes valuable insights into the predictors of sleep disorders and paves the way for further research and practical applications in improving sleep health outcomes.

- ## ACKNOWLEDGMENT :

“It is not possible to complete a project without the assistance and encouragement of other people.this one is certainly not the exception.” On the very outset of this project,I would like to extend my sincere and heartfelt obligation towards all the personages who have helped me a lot in this endeavor.Without their active guidance,help,cooperation and encouragement,I would not have made headway in this project.
I am ineffably indebted to our head of the statistics department, Sir Debesh Roy and Sir.Arup Kumar Hait for conscientious guidance and encouragement to accompalish this assignment.
I am extremely thankful and pay my gratitude to my faculty Sir Kiranmoy chatterjee, and Sir Soumyadeep das for their valuable guidance and support on completion of this project in its presently.
I extend my gratitude to my college Bidhannagar Govt. College for giving me this opportunity.
I also acknowledge with a deep sense of reverence,my gratitude towards my parents who have always supported me morally as well as economically.
At last but not the least gratitude goes to all of my friends who directly or indirectly helped me to complete this project.

- REFERENCE :

(1) An Introduction to Categorical Data Analysis by Alan Agresti (z-lib.org)
(2) (Wiley series in probability and statistics) Alan Agresti - Categorical data analysis-Wiley (2013)
(3) Kaggle.com
(4)https://en.wikipedia.org/wiki/Multinomial_logistic_regression